

Is it time for ethical boards in top-tier machine-learning conferences? A case-study on an Interspeech challenge about children suicidability.

Vincent P. Martin^{1,*†}

¹Université de Lorraine, CNRS, Inria, LORIA, Nancy, France

Abstract

Automatic speech analysis is presented as a promising tool for mental health monitoring, including in sensitive contexts such as suicide-risk detection. In 2025, the SpeechWellness Challenge at Interspeech invited participants to develop machine-learning systems to classify suicidal risk in children and adolescents from speech recordings. Drawing on my participation in this challenge, this opinion paper examines its scientific and ethical foundations as a case study for the governance of clinical machine-learning research. Using a five-step framework for ethical machine learning in health care, I analyze the challenge design, dataset, clinical rationale, and evaluation procedures. I argue that, despite addressing an important and understudied issue, the challenge raises substantial concerns: the clinical utility of binary suicide-risk prediction from short speech recordings remains unclear; the data and labels provide limited insight into real-world psychiatric assessment; the evaluation setting favors technical competition over clinically meaningful validation; and post-deployment implications are insufficiently addressed. This case study suggests that top-tier speech and machine-learning conferences need stronger mechanisms to evaluate the ethical and clinical implications of submitted work, especially when vulnerable populations and mental-health outcomes are involved. I conclude by advocating for stronger interdisciplinary collaboration with clinicians and dedicated ethical review boards, with clearer reporting requirements.

Keywords

Clinical AI ethics, Speech biomarkers, Suicide-risk detection, Machine-learning challenges

1. Introduction

1.1. Context.

Automatic speech analysis is usually presented as a promising biomarker for psychiatry [1]. Usually framed under the terms *speech biomarkers* or *vocal biomarkers*, medical applications represent a growing interest of the speech processing community [2].

Speech biomarkers are an interdisciplinary field at the crossroads of medical sciences, speech science, and computer science, with publications spanning from specialized biomedical research journals to computer science and signal processing conferences. In the latter field, challenges are regularly organized within Interspeech, one of the major scientific conferences in the field. Usually, they consist of providing participants with a consolidated dataset and asking them to develop a machine learning pipeline to estimate a health dimension (e.g. the medication status of Parkinson patients [3] or the 5-year prognosis of Alzheimer patients [4]). Unlike *workshops*, the papers submitted to these *special sessions* are published in the main proceedings of the conference, alongside regular papers.

At the end of 2024, an Interspeech challenge caught our attention: the SpeechWellness challenge [5], aiming at automatically detecting suicidal risk in children and adolescents. I participated in this challenge, seeing the opportunity to publish in a top-tier conference a new approach to speech analysis, with an optimistic acceptance rate on such challenges¹. Even after the participation of my team, the

First Workshop on Ethics in Clinical AI Applications (ETHICAIA), co-located with the 35th International Joint Conference on Artificial Intelligence (IJCAI-ECAI 2026), Bremen, Germany

✉ vpmartin@inria.fr (V.P. Martin)

🌐 <https://members.loria.fr/VMartin/> (V.P. Martin)

🆔 0000-0001-5937-4259 (V.P. Martin)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹Challenges are known to have high acceptance rates on new / innovative approaches, even if they do not necessarily obtain the best performance. For instance, our article was accepted with a Weak Accept to counterbalance a Weak Reject and a

clinical rationale for such a challenge remains blurry, and its clinical utility is uncertain. Worse, during my participation in this challenge, multiple ethical breaches were observed.

A useful resource to investigate the ethical aspects of such a challenge is the five-step framework proposed by Chen et al. in their reference article [6]: i) problem selection; ii) data collection; iii) outcome definition; iv) algorithm development; v) post-deployment considerations.

1.2. Objectives.

In this opinion paper, I draw from the five steps identified by Chen et al. to investigate the ethical aspects of the SpeechWellness challenge. This paper is structured as follows. In Section 2, I describe in detail the challenge, the provided data, and I report the performance of the challenge participants. Then, in Section 3, rooted in my experience of this challenge, I discuss, step by step, the ethical aspects – being positive or negative – of the challenge. Finally, in Section 4, I conclude on the need for ethical boards and strong interdisciplinarity for research crossing machine learning development and clinical applications.

2. Description of the challenge

2.1. General description

The SpeechWellness challenge [5] was organized within the international conference Interspeech 2025. After the deadline for registration on 15 January 2025, the participants had less than one month to design their pipeline, submit up to three predictions on the unlabelled test set, and write a full paper for Interspeech.

At the time of the challenge, the resources for the participants were the dataset (cf. below), the preprint of the article introducing the challenge², a GitHub repository containing the code for the proposed baselines³ and a website serving as the main information broadcast to all participants⁴.

2.2. Dataset

The SpeechWellness dataset contained the recordings of 600 Chinese children and adolescents (aged 10-18) from 47 schools in Yunfu, southern China. They were recorded performing three tasks: i) answering a spontaneous question, “Have you ever experienced moments of extreme emotional distress? How do you manage such feelings?” (total duration: 5h04min); ii) reading the text *The North Wind and The Sun* (total duration: 6h34min); iii) description of a photo featuring a face with a pronounced expression (total duration: 3h16min). Informed consent was obtained from all the participants and their guardians, and the participants were informed about their right to withdraw or decline at any time. The interviews were conducted by trained interviewers, and data collection was performed under the supervision of psychiatrists. All the recordings were made in Mandarin. Their suicide risk was assessed using the MINI-KID, a standardized interview guide comprising 8 questions, from which a suicidality risk score can be derived (between 1 and 36). It is reported in Table 1.

Among the 600 participants, 300 (50%) were labelled as being at suicidal risk. The training (n=400), validation (n=100) and test (n=100) sets were provided by the organizers. They had the same ratio of positive samples (50%), the same ratio between men and women (30%/70%) and the same age distribution (mean age around 14). No further information was given about the data, but the challenge data are strongly suspected to be a subset of a larger corpus collected by the same authors (cf. Section 3.4).

Reject recommendation from the three reviewers. While we still think that this paper is of interest for the community, the gap between reviewers' recommendations and the meta-reviewer's decision is intriguing.

²<https://arxiv.org/pdf/2501.06474v1>

³https://github.com/speechwellness/SpeechWellness-1_Baseline

⁴<https://speechwellness.github.io/>, consulted on 21/05/2026

Question	No	Yes	Points
In the past month, have you wished you were dead?	☐	☐	1
Have you thought about harming yourself?	☐	☐	2
Have you had suicidal thoughts?	☐	☐	4
Have you thought about a suicide method?	☐	☐	1
Have you intended to act on suicidal thoughts?	☐	☐	2
Have you made a suicide attempt?	☐	☐	6
Have you ever attempted suicide in your lifetime?	☐	☐	10

Risk Severity Scoring

Score Range	Risk Level
1–8	Low Risk
9–16	Moderate Risk
17+	High Risk

Table 1

MINI-KID for suicidal risk assessment.

System	Ac.	Lex.	LLM	Dev.	Test
Baseline					
Baseline 1	✓			53%	51%
Baseline 2		✓		56%	61%
W.o. anonymization					
Baseline 1	✓			?	?
Baseline 2		✓		58%	60%
Participants					
[8]	✓	✓	✓	77%	74%
[9]			✓	67%	68%
[10]	✓	✓		63%	63%
[11]	✓	✓		69%	61%
[12]	✓			63%	55%
[13]			✓	68%	54%

Table 2

Performances and nature of the features or embeddings used by the challenge participants. Ac.: Acoustic; Lex.: Lexical

2.3. Baselines

Two baselines were proposed for the challenge. The first one relies on the standard eGeMAPS acoustic feature set [7] and a Support Vector Machine (SVM). This approach (*Baseline 1*), achieved accuracies of 53% on the development set and 51% on the test set. The second approach (*Baseline 2*), consists of the association of wav2vec2 + BERT, achieving a performance of 56% on the development set and 61% on the test set.

2.4. Participants

While the authors of the challenge report 127 registrations [5] and 186 system submissions, we only retrieved six papers corresponding to participation in the challenge in the proceedings of the Interspeech 2025 conference (4.7% of registered teams). Their performances and the nature of their features are reported in Table 2.

3. Ethical framework

This section investigates the ethical consideration of the SpeechWellness challenge under the five steps outlined in the reference article of Chen et al. [6].

3.1. Step n°1 - Problem selection

No previous work on this topic. We acknowledge that the organizers of the challenge have addressed an understudied topic in the speech community. Indeed, the organizers of the challenge propose to detect suicidal risk, which is largely underrepresented among the psychiatric manifestations studied in speech conferences or journals⁵: such articles usually focus on depression [1], with emblematic corpora such as the DAIC-Woz [15]. Moreover, the authors propose to estimate suicidal risk in children and adolescents, a particularly vulnerable population for which suicide is the third leading cause of death worldwide in the 15-19 age range [16]. To our knowledge, no previous article has proposed designing speech biomarkers related to the mental health of this particular population. The task seems useful at first glance: estimating instantaneous suicide risk using speech could allow it to be implemented inside smartphone apps, promising early intervention, before a potential suicide attempt.

Suicide is not an emotion. Far from being an emotional state, suicidal risk is a multidimensional, longitudinal, and dynamic clinical construct. Despite this evidence, the challenge itself is named *SpeechWellness*, and, in the conference and program, is presented in a session named *Emotional Distress in Speech*. This confusion is deeply rooted in the idea that suicide is the acme of mental distress, resulting from a gradually increasing ill-being. While this is the case for some patients who proceed to a suicide attempt [16], some patients perform suicide attempts without depression or a known history of anxiety [17]. Suicide risk – all the more in children and adolescents – needs to be treated with the gravity it deserves: suicide is not ‘emotional distress’, and needs to be called as such⁶.

Why do researchers focus on SER? Speech Emotion Recognition (SER) is an emblematic task of the speech community, for which credible performance has been obtained for decades now [19]. This task played a significant and normative role in shaping the speech processing community, and in particular the field of vocal biomarkers, through its challenges, datasets (e.g. emo-DB), and feature-extraction software (e.g. eGeMAPS). Unfortunately, because SER has become such a central reference point, more complex problems, such as mental health, are often reframed as familiar emotion-recognition tasks based on speech signals.

Depression is not an emotion. For instance, articles may confound negative emotions with depression. In that regard, experimental results can be misleading: in some cases systems trained on emotion labels may achieve good performance in detecting depression [20]. However, these good performance are more likely the results of a sampling bias of the depression corpus (i.e. contrasting depressive participants with negative emotions vs. healthy participants with positive emotions) and are not proof that the system has indeed generalized depressive patterns in speech from emotion datasets. Similarly to suicide, depression is not an emotion⁷: it is a chronic mood disorder with potentially serious adverse effects, functional impairment, comorbidities. This complexity contrasts with systems that aim to estimate it using only one speech sample associated with sad emotion, whereas it is a chronic condition. Participants with such a disorder need structured treatment and follow-up rather than gadgets estimating their emotion from speech.

Conclusion n°1. Suicidal risk in children and adolescents is an understudied problem: by putting it under the spotlights, the authors of the challenge contribute to making it visible to the speech processing community. However, there are still problems in how this task is presented, by the authors and the community (through Interspeech program committee): SER systems are to be left to researchers

⁵Interestingly, for LLMs, it is one of the most represented studied tasks related to mental health [14]

⁶The explicit naming of *suicide* in general media – under strict conditions – has even been associated with beneficial effects, under the Papageno effect [18]

⁷In the same vein, in a recent article, we have deconstructed the preconception that in emergency call centers, the emotion of the caller would be linked to the degree of emergency of the situation [21].

working on emotion-related issues, and the other problems are to be tackled for what they are: suicide and depression are not emotions.

3.2. Step n°2 - Data collection

Speech recordings. Regarding data collection, the authors of the challenge took care to record all the participants using the same recording devices, in similar recording environment [5], which avoids spurious correlations between the label and recording equipment or environmental noise [22]. While being basic knowledge for any speech scientist, during the Covid-19 era, some datasets were recorded with different setups in different locations (e.g. patients in hospitals and healthy controls at their home), artificially inflating the performances of respiratory infection detection [23].

Urban vs. rural population. The recordings were performed in schools of the Chinese city of Yunfu, which counted more than 3 million inhabitants in 2025⁸: this focus only covers children and adolescents living in urban areas or on the outskirts of urban areas. While, in China, schooling is free and compulsory, this focus overlooks rural students, who come from families with lower cultural capital and who generally attend schools closer to their homes, which have fewer resources, leading to less prestigious educational pathways [24, 25]. Complementary studies including rural schools would allow comparisons of the effect of the living environment on suicidability in children and adolescents, but also allow for Mandarin-speaking proficiency comparison between the two sub-populations.

Data sharing. A major ethical breach I encountered during the challenge was the way the organizers shared the data with the challenge participants. Indeed, the organizers seem to care about the security of the file transfer: not only they anonymized the voice of the participants (cf. Section 3.4), but they also separately encrypted the training, development, and test sets of their data in .zip files with 8-character passwords (mixing uppercase letters, lowercase letters, numbers and special characters). However, despite these precautions, they still demonstrated a high level of amateurism by hosting them in a Google Drive folder⁹ and sending the three passwords (for training, validation, and test sets) in a single email using the same Gmail address as the Google Drive.

Whereas this case is notable for the contrast between the authors' good intentions and their actual practices, the way this dataset is stored and shared is no worse than the DAIC-WoZ, which is not encrypted and downloadable by accessing a simple URL¹⁰; or the Android corpus, which is hosted without any data encryption on a Dropbox folder¹¹.

Their accessibility is a key reason for their success and their (over)use in the field of vocal biomarkers of depression: independently of their scientific rationale, the fact that they can be downloaded with a few clicks makes them easily accessible resources for researchers of the field. However, this ease is to be put into balance with ethical risks of reidentification, and all the adverse effects for participants who have their personal health data publicly available [26, 27].

Conclusion n°2. The authors of the challenge took credible precautions in the methodology for speech recordings. However, comparative data with rural schools are lacking and would allow comparisons of both suicidal risk rates and Mandarin-speaking proficiency comparison between the two sub-populations. Moreover, the ease of access to data should comply with robust security and ethical standards, guaranteeing that participants do not suffer from potential adverse consequences of reidentification. Finally, from an ethical point of view, it is unclear if children and adolescents who presented suicidal risks were offered care or not.

⁸<https://www.yunfu.gov.cn/>

⁹Files are still accessible on the same link, as of 24 May 2026

¹⁰I do not put it here, but it is really not far from the root website introducing the dataset: <https://dcapswoz.ict.usc.edu/>

¹¹Direct link available on <https://github.com/androidscorpus/data>

3.3. Step n°3 - Outcome definition

Binary classification (again). The challenges relied on the estimation of a binary variable, i.e. whether the participants had ‘suicidal risk’ or not. While it was not detailed by the authors of the challenge, we hypothesize that the MINI-KID was filled by the interviewer: contrary to usual questionnaires (e.g. the PHQ-9 used in the DAIC-WoZ), the MINI-KID is more a structured interview guide rather than a psychometric tool. Similarly, we can only hypothesize that the binary labels were derived from having a MINI-KID score greater than or equal to 1 (cf. Table 1).

Binary classification is usually chosen to simplify machine learning problems, by easing the convergence of models (as compared to multiclass problems). However, this artificially delimits two heterogeneous groups, for which it is then difficult to find the relevant border in the multidimensional feature space: instead of providing only the binary label, the organizers should have provided the raw MINI-KID scores, which would eventually have allowed the identification of voice patterns linked to the severity of suicidal risk – and potentially to better distinguish them.

Other symptoms expressed through voice. Worse still, suicidal risk is linked to numerous comorbid conditions [28], themselves manifesting in speech, such as low mood [29], depression [30], fatigue [31], stress [32], or sleep disorders [33], among others. While the authors balanced their training, development, and test subsets regarding sex, age and suicidability status, they did not measure these dimensions (or at least, they did not provide them to the participants of the challenge): spurious correlations within the data may prevent the generalizability of systems trying to automatically detect suicidal risk independently of these comorbid symptoms, or, on the contrary, would allow good classification performances based on a related symptom (sampling bias) instead of suicidability itself. This is in line with a previous article arguing for the labelling of datasets with multiple symptoms [34], from which the disorders are identified, instead of binary disorder labels (e.g. ‘Depressive vs. Not depressive’).

Conclusion n°3. Instead of providing only the binary label for the presence of suicidal risk, the organizers should have i) (costless) provided the raw scores to the MINI-KID or ii) (better) assessed and provided the most prevalent related symptoms in addition to the suicidality label.

3.4. Step n°4 - Algorithm development

Performance and accuracy. In the challenge, the reference metric (the one for the leaderboard) was accuracy, defined by the ratio of correctly identified samples. However, when submitting for an evaluation, the organizers of the challenge returned the confusion matrix of the evaluation on the blind test set, allowing computation of these different metrics¹².

While accuracy is a relevant metric given the balancing of the classes in each subset, the challenge could have emphasized the need to measure and control False Negatives (e.g. through Recall) and False Positives (e.g. through Precision). Indeed, in real-world implementations, the two cases have different yet crucial negative consequences. On the one hand, False Negatives are associated with missed suicidal risk in participants who may need care to prevent suicide attempts. On the other hand, False Positives may lead to unnecessary preventive interventions, wasting healthcare professionals’ time and resources. In the most severe cases, they may also result in coercive protective measures, including enforced hospitalization, depriving patients of their liberty.

Group fairness and other evaluations. While the sex and the age of the participants were provided, no system to our knowledge evaluated fairness. In our view, this mainly comes from the brevity of the Interspeech papers (only 4 pages, double columns, 1 additional page for references in 2025), but also – and mostly – from the drive of the challenge, which is classification performance. It doesn’t matter if the system is biased or has consumed a ridiculously high amount of energy; in challenges, nothing is too expensive when it comes to achieving the highest performance.

¹²Knowing that the test set is balanced could have been used by dishonest participants for the inference on the test set.

Anonymization process. In order to be able to share their sensitive data (which is both personally identifiable and contains health-related information), the challenge organizers used neural voice conversion techniques “to alter the timbre of the participants’ voices while preserving prosody and rhythm” to anonymize the participants’ audio recordings [5]. They measured the efficacy of their anonymization procedure using Equal Error Rate and assert that they kept useful information by preserving Character Error Rate (CER) in Automatic Speech Recognition (ASR), which is on par with the standards of the speech anonymization community [35].

Despite these efforts, I still hypothesize that the application of such an algorithm destroyed relevant acoustic information for the identification of suicidal risk. Indeed, the measurement of utility by a low CER on ASR only testifies to the conservation of the lexical content, but does not provide any argument in favour of the conservation of the acoustic properties of speech that could be useful for the detection of suicidal risk. Worse, the organizers state that, using their anonymization procedure, “the essential features for developing the automatic suicide risk detection system remain intact.” without first identifying what these ‘essential features’ might be, and contradicting previous observations [30].

Furthermore, the organizers of the challenge even anticipated this criticism by reporting performance for their Baseline 2 before anonymization (cf. Table 2), stating that this “suggests that the anonymization process preserves the information relevant to the task”. However, they did not report such performance for Baseline 1, based on acoustic characteristics of speech, and are thus not able to support such a statement. Conversely, all systems based only on acoustic characteristics showed test performances lower than or equal to 55% (Table 2), pointing to a tendency for the anonymization process to destroy useful acoustic information.

Finally, last but not least, the authors did not report the potential bias that could have been introduced by their anonymization procedure [36]. Does it work similarly for each subgroup (sex, age, suicidal risk)?

Is it a baseline or not? The official baseline of the challenge (Baseline 1) is extremely low (53% accuracy on the development set, 51% on the test set). Surprisingly, the authors also propose a second baseline, that we named *Baseline 2* but which they call ‘Baseline-bonus’: although still very low, this second baseline outperforms Baseline 1 by 10% on the test set. However, in order to attract participants with an easily surpassable baseline, Baseline 1 is promoted as the reference to beat for the challenge – in other words, the objective is to do better than a coin flip (50% accuracy).

However, the challenge organizers do not provide sufficient transparency in the way they present these baselines. Indeed, the previous year, in the Interspeech 2024 proceedings, the same authors (although in a different order) published a paper reporting accuracy and F1 scores higher than 80% for estimating suicidal risk in children and adolescents, based on a corpus with the same characteristics as the one proposed for the challenge [37]. Three differences should nevertheless be noted between the challenge corpus and the one used in 2024. First, the dataset used in 2024 contains recordings from 1,179 participants: we therefore assume that the dataset proposed for the challenge is a subset of this corpus, designed by the organizers to be artificially balanced with respect to suicidal risk. Second, this performance was obtained on a task that was not included in the challenge: participants had to answer the question “How would you describe yourself?”. It seems that lexical analysis of the response to this task could be more discriminative than the tasks included in the challenge. Finally, these data did not undergo the anonymization process applied to the challenge data, suggesting that their acoustic characteristics, in particular timbre, may be exploited for suicidal-risk identification.

Similarly, the pipelines proposed as baselines are far less complex than the pipeline used by the authors in their 2024 paper, notably because the latter concatenated acoustic and lexical embeddings, which they did not propose as a baseline for their challenge.

Conclusion n°4. Challenges need to go further than just promoting the systems with the highest accuracies: in addition to high discrimination capabilities, they should promote other aspects of pipeline evaluation, such as low energy consumption, or absence or mitigation of biases. Moreover, while the

anonymization of speech may be used to protect the data from reidentification, the impact on the downstream task (suicide risk identification) has to be further studied. Finally, challenges should include baselines on par with the state of the art of the studied task, and should not artificially deflate the performances of such baselines to attract participants, for whom the task of beating the baseline is thus made easier.

3.5. Step n°5 - Postdeployment considerations

Model documentation. None of the participants shared their code online, or provided complete model documentation. Again, in our view, this comes from the desire of some participants to keep their code undisclosed for commercial reasons (e.g. industrial participants), but also from constraints of the conference itself (4-page papers, strict deadline) and from the very nature of challenges in general, which tend to favor ‘quick and dirty’ approaches to the detriment of mature, slow science [38].

From labs to clinical practice. The performances achieved by the challenge participants remain very modest, and their use in real-world conditions seems unlikely. By contrast, on similar data, the authors managed to obtain performances exceeding 80% accuracy and F1-score. However, strong classification scores alone do not imply that such systems would meaningfully “move the needle in reducing suicide, reducing hospitalizations, improving recovery” (Thomas Insel, former NIMH director, 2017). Indeed, despite more than 15 years of reported systems that appear to ‘work’ in experimental settings [30], suicide-risk prevention based on automatic voice analysis has still not been implemented in real-world care. This is due in particular to the lack of clinical validity of these devices, which rarely make the effort to comply with the requirements of medical standards [39], but also to the diversity of epistemic communities, which continuously renegotiate what a digital biomarker is in psychiatry [40].

Ethical considerations. Beyond these concerns, the organizers also undermined the possibility of any real-world validation of the task’s feasibility. Indeed, in the preprint and on the website presenting the challenge, the only ethical risks raised by the organizers focused on reidentification risks. It was only one week before the paper submission deadline that the organizers formally asked participants to include the following paragraph in their paper (and included it themselves in the final version of their article):

“The findings presented in this study are based on the scoring framework of the MINI-KID scale, which assesses current suicide risk as at risk or no risk. This classification reflects participants’ immediate responses to the MINI-KID assessment and should not be interpreted as a prediction of future suicidal behavior. Although the MINI-KID suicide module is widely recognized as a gold standard for assessing current suicide risk among adolescents, it has limitations. It relies heavily on self-reported data, which may lead to underreporting or misinterpretation of symptoms, and its fixed set of items may not fully capture the complex and dynamic nature of suicidal ideation and behavior. Accordingly, the results reported herein are strictly confined to the context of this assessment.”

In other words, fearing that their system might actually be deployed to predict suicidal behavior, the challenge organizers highlight the limitations of the model (it is based only on a limited set of items and ‘subjective’ data) and emphasize the local nature of the results obtained (which are valid only within the context of the challenge). By stating that the results obtained through the challenge are limited to it, they hope this statement “will safeguard the scientific accuracy and ethical integrity of the challenge while enhancing inclusiveness and rigor, which are central to our shared goals.” (Conclusion of the email sent to all participants).

If the results obtained on this challenge are limited to it, then what are the benefits for patients?

Conclusion n°5. Structuring the task as a challenge makes it difficult both to produce mature, well-documented, and properly validated code or models, and to move toward the real-world implementation of such systems, especially in view of the requirements of biomedical research. Here again, a careful balance must be found between ethical considerations and scientific progress, so as not to fall into excessive precaution that would invalidate any scientific result.

4. Conclusion

Although promising and addressing a topic that remains largely under-studied within the speech-biomarker community, this challenge ultimately appears to have had little concrete impact.

Who benefits from this challenge? Despite the organizers' enthusiasm in highlighting the 68.8% of submitted systems surpassing the Baseline (random) and the 14 systems (7.5%) that achieve better performance than the Baseline-bonus accuracy of 61% on the test set, we do not see who really benefits from this kind of challenge. *Clinicians and patients?* We do not think so, since the richness of the interaction between patients and clinicians is reduced to a voice recording used to estimate a binary variable. Moreover, neither clinicians nor patients will probably ever see the systems designed for the challenge, which will likely gather dust like many other systems developed in such challenges. *The challenge participants?* Whether the goal is to access a dataset or to promote new machine-learning techniques, the data are not of sufficient quality (cf. previous discussions), and the performances achieved make it difficult to promote new methods as truly disruptive. *The challenge organizers, then?* As a reward for their efforts in preparing the data, managing the systems to be evaluated, designing the baselines, and carrying out all the infrastructural work that made this challenge possible, the organizers benefit from an additional paper which, thanks to the challenge, will have been cited six times (with a seventh citation in the present article).

While such an interdisciplinary topic is inevitably exposed to scientific risks, we argue that many of the limitations identified in this article could have been anticipated by mobilizing two resources.

Interdisciplinary collaborations. First, the challenge could have relied on stronger interdisciplinary collaboration with child psychiatrists and psychiatrists specializing in suicide-risk management. By this, we mean strong interdisciplinarity rather than mere multidisciplinary: the point is not only to hold a few meetings with psychiatrists to design the study, but rather to engage with clinical practice in the field, and to become familiar with the clinical practice, scientific foundations, and clinical realities of child psychiatry. In return, such field engagement may also help psychiatrists become more familiar with machine learning and data engineering, fostering constructive exchanges that make datasets relevant and robust according to the expectations of both communities [40].

Ethical review committee. A second resource that could help to put Interspeech on the right track would be the creation of an ethics review committee, similar to what exists in the NLP community through the ACL Rolling Review (ARR) system: each scientific reviewer has the opportunity to flag the article they are reviewing as requiring an ethical review¹³. The submission is then scrutinized by ethical reviewers dedicated to these aspects, under the supervision of ethics chairs: independently of the scientific quality of the content, articles may risk rejection based on this ethical scrutiny [41]. However, implementing such a system would also require authors to be able to detail these aspects in their papers, for instance in a dedicated paragraph that does not count toward the maximum length accepted by the conference, which is still not the case for Interspeech.

Although imperfect, notably because it requires recruiting such ethics reviewers in a context where recruiting reviewers is already difficult, this solution could drastically improve the ethical aspects of Interspeech papers dealing with health-related topics. It could also remove the loophole that allows

¹³<https://aclrollingreview.org/ethics-flagging-guidelines/>

teams promoting unethical practices, which would not be publishable in biomedical research journals because these venues systematically require such ethical information, to publish in a well-ranked conference instead.

Conclusion. While the reflections developed in this article are based on a single challenge organized at Interspeech, previous concerns about challenges and AI ethics have been raised in the largest NLP [42] and machine-learning conferences [43, 44]. Ultimately, only strong interdisciplinarity and sustained attention to ethical issues will allow vocal biomarkers – and, more broadly, clinical machine-learning systems – to keep their promises.

Acknowledgments

VPM has received financial support from the French Research Agency ANR through the axis “Autonom-Health” of the PEPR “Santé Numérique” (Grant agreement n°ANR-22-PESN-0009), and from the InEx-tenso project (ANR-23-IAS1-0004).

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

- [1] D. M. Low, K. H. Bentley, S. S. Ghosh, Automated assessment of psychiatric disorders using speech: A systematic review, *Laryngoscope Investigative Otolaryngology* 5 (2020) 96–116. doi:10.1002/liv.2.354.
- [2] J. Robin, J. E. Harrison, L. D. Kaufman, F. Rudzicz, W. Simpson, M. Yancheva, Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations, *Digit Biomark* 4 (2020) 99–108. doi:10.1159/000510820.
- [3] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, F. Weninger, The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson’s & eating condition, in: *Interspeech 2015*, 2015, pp. 478–482. doi:10.21437/Interspeech.2015-179.
- [4] S. Luz, F. Haider, S. de la Fuente, D. Fromm, B. MacWhinney, Alzheimer’s Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge, in: *Interspeech 2020*, 2020, pp. 2172–2176. doi:10.21437/Interspeech.2020-2571.
- [5] W. Wu, Z. Cui, C. Lei, Y. Duan, D. Qu, J. Wu, B. Zhou, R. Chen, C. Zhang, The 1st SpeechWellness Challenge: Detecting Suicide Risk Among Adolescents, in: *Interspeech 2025*, 2025, pp. 399–403. doi:10.21437/Interspeech.2025-124.
- [6] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, M. Ghassemi, Ethical Machine Learning in Healthcare, *Annu Rev Biomed Data Sci* 4 (2021) 123–144. doi:10.1146/annurev-biodatasci-092820-114757.
- [7] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, K. P. Truong, The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing, *IEEE Trans. Affective Comput.* 7 (2016) 190–202. doi:10.1109/TAFFC.2015.2457417.
- [8] Y. Gao, J. Fu, L. Guo, H. Liu, Leveraging Large Language Models for Spontaneous Speech-Based Suicide Risk Detection, in: *Interspeech 2025*, 2025, pp. 404–408. doi:10.21437/Interspeech.2025-1483.
- [9] F. Roquefort, A. Ducorroy, R. Riad, In-context learning capabilities of Large Language Models to detect suicide risk among adolescents from speech transcripts, in: *Interspeech 2025*, 2025, pp. 414–418. doi:10.21437/Interspeech.2025-2083.

- [10] X. Chen, R. Yu, Y. Tan, Y. Li, Q. Qian, Y. Lin, Predicting Adolescent Suicidal Risk from Multi-task-based Speech: An Ensemble Learning Approach, in: *Interspeech 2025*, 2025, pp. 409–413. doi:10.21437/Interspeech.2025-2419.
- [11] J. Krzywdziak, B. Eljasiak, J. Stępień, M. Świątek, A. Pruszek, Leveraging Text and Speech Processing for Suicide Risk Classification in Chinese Adolescents, in: *Interspeech 2025*, 2025, pp. 394–398. doi:10.21437/Interspeech.2025-2755.
- [12] V. P. Martin, C. Brazier, M. Amblard, M. Musiol, J.-L. Rouas, Network of acoustic characteristics for the automatic detection of suicide risk from speech. Contribution to the 2025 SpeechWellness challenge by the Semawave team, in: *Interspeech 2025*, 2025, pp. 424–428. doi:10.21437/Interspeech.2025-599.
- [13] J.-W. Kim, W. Oh, H. Yoon, S.-H. Yoon, D.-J. Kim, D.-H. Lee, S.-Y. Lee, C.-M. Yang, Language-Agnostic Suicidal Risk Detection Using Large Language Models, in: *Interspeech 2025*, 2025, pp. 419–423. doi:10.21437/Interspeech.2025-1372.
- [14] C. Bleuze, K. Fort, V. P. Martin, A. Névél, Large Language Models for Mental Health Prediction: A Scoping Review of Bias and Clinical Utility Documentation in 2019-2024, Under press for *JMIR AI* (2026).
- [15] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, L.-P. Morency, The Distress Analysis Interview Corpus of human and computer interviews, in: *LREC 2014*, 2014, pp. 3123–3128.
- [16] D. Wasserman, V. Carli, M. Iosue, A. Javed, H. Herrman, Suicide prevention in childhood and adolescence: A narrative review of current knowledge on risk and protective factors and effectiveness of interventions, *Asia-Pacific Psychiatry* 13 (2021) e12452. doi:10.1111/appy.12452.
- [17] R. P. Auerbach, J. G. Stewart, S. L. Johnson, Impulsivity and Suicidality in Adolescent Inpatients, *Journal of Abnormal Child Psychology* 45 (2017) 91–103. doi:10.1007/s10802-016-0146-8.
- [18] J. Domaradzki, The Werther Effect, the Papageno Effect or No Effect? A Literature Review, *IJERPH* 18 (2021) 2396. doi:10.3390/ijerph18052396.
- [19] A. Hashem, M. Arif, M. Alghamdi, Speech emotion recognition approaches: A systematic review, *Speech Communication* 154 (2023) 102974. doi:10.1016/j.specom.2023.102974.
- [20] Y. Shen, H. Yang, L. Lin, Automatic Depression Detection: An Emotional Audio-Textual Corpus and A Gru/Bilstm-Based Model, in: *ICASSP 2022*, 2022, pp. 6247–6251. doi:10.1109/ICASSP43922.2022.9746569.
- [21] E. Stasica, C. Joly, A. Lecomte, V. P. Martin, R. Serizel, E. Vincent, T. Chouihed, Revisiting Emotion-Based Triage: Evidence from French Emergency Call Data, in: Under press for *Interspeech 2026*, 2026.
- [22] L. Gauder, P. Riera, A. Slachevsky, G. Forno, A. M. García, L. Ferrer, A Toolkit for Detecting Spurious Correlations in Speech Datasets, [arxiv preprint] (2026). doi:10.48550/ARXIV.2604.26676.
- [23] H. Coppock, L. Jones, I. Kiskin, B. Schuller, COVID-19 detection from audio: Seven grains of salt, *The Lancet Digital Health* 3 (2021) e537–e538. doi:10.1016/S2589-7500(21)00141-2.
- [24] Q. Fu, Q. Ren, Educational Inequality under China’s Rural–Urban Divide: The *Hukou* System and Return to Education, *Environ Plan A* 42 (2010) 592–610. doi:10.1068/a42101.
- [25] Y. Hu, Family or school? Understanding the urban–rural achievement gap and its generating mechanisms in China, *Asia Pacific Educ. Rev.* 26 (2025) 1055–1070. doi:10.1007/s12564-025-10066-x.
- [26] W. N. Price, I. G. Cohen, Privacy in the age of medical big data, *Nature Medicine* 25 (2019) 37–43. doi:10.1038/s41591-018-0272-7.
- [27] E. Ford, S. Pillinger, R. Stewart, K. Jones, A. Roberts, A. Casey, K. Goddard, G. Nenadic, What is the patient re-identification risk from using de-identified clinical free text data for health research?, *AI and Ethics* 5 (2025) 4441–4454. doi:10.1007/s43681-025-00681-0.
- [28] A. Junus, P. S. F. Yip, Preventing comorbidity between distress and suicidality: A network analysis, *npj Mental Health Res* 2 (2023) 2. doi:10.1038/s44184-023-00022-1.
- [29] T. James, T. M. Reddy Yenumula, T. Mergist, R. Muhammad, A. Harimi, K. Khalil, A. Kumar, Speech Emotion Recognition Using ML Models and Audio Features, in: *2025 IEEE International Systems*

- Conference (SysCon), 2025, pp. 1–7. doi:10.1109/SysCon64521.2025.11014858.
- [30] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T. F. Quatieri, A review of depression and suicide risk assessment using speech analysis, *Speech Communication* 71 (2015) 10–49. doi:10.1016/j.specom.2015.03.004.
 - [31] H. P. Greeley, E. Friets, J. P. Wilson, S. Raghavan, J. Picone, J. Berg, Detecting Fatigue From Voice Using Speech Recognition, in: *IEEE International Symposium on Signal Processing and Information Technology*, 2006, pp. 567–571. doi:10.1109/ISSPIT.2006.270865.
 - [32] M. Kappen, J. van der Donckt, G. Vanhollebeke, J. Allaert, V. Degraeve, N. Madhu, S. Van Hoecke, M.-A. Vanderhasselt, Acoustic speech features in social comparison: How stress impacts the way you sound, *Sci Rep* 12 (2022) 22022. doi:10.1038/s41598-022-26375-9.
 - [33] V. P. Martin, J.-L. Rouas, P. Philip, Automatic detection of sleepiness-related symptoms and syndromes using voice and speech biomarkers, *Biomedical Signal Processing and Control* 91 (2024) 105989. doi:10.1016/j.bspc.2024.105989.
 - [34] V. P. Martin, J.-L. Rouas, Estimating Symptoms and Clinical Signs Instead of Disorders: The Path Toward The Clinical Use of Voice and Speech Biomarkers In Psychiatry, in: *ICASSP 2024*, 2024, pp. 10606–10610. doi:10.1109/ICASSP48485.2024.10445888.
 - [35] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, M. Todisco, The VoicePrivacy 2024 Challenge Evaluation Plan, [Technical report] (2024). doi:10.48550/ARXIV.2404.02677.
 - [36] A. Leschanowsky, Ü. E. Gaznepoglu, N. Peters, Voice Anonymization for All-Bias Evaluation of the Voice Privacy Challenge Baseline Systems, in: *ICASSP 2024*, 2024, pp. 4785–4789. doi:10.1109/ICASSP48485.2024.10447137.
 - [37] Z. Cui, C. Lei, W. Wu, Y. Duan, D. Qu, J. Wu, R. Chen, C. Zhang, Spontaneous Speech-Based Suicide Risk Detection Using Whisper and Large Language Models, in: *Interspeech 2024*, 2024, pp. 2915–2919. doi:10.21437/Interspeech.2024-1895.
 - [38] L. Alleva, Taking time to savour the rewards of slow science, *Nature* 443 (2006) 271–271. doi:10.1038/443271e.
 - [39] Show us the evidence for the value of medical AI, *Nat Med* 32 (2026) 1163–1163. doi:10.1038/s41591-026-04389-4.
 - [40] V. P. Martin, E. Gratreau, S. Mouchabac, G. Serrano, J.-A. Micoulaud Franchi, Digital psychiatric biomarkers beyond the walls: A Science and Technology Studies view, [Under review] (2026).
 - [41] A. Rossetto, A. Midelet, S. Baillieul, R. Tamisier, J.-C. Borel, A. Prigent, S. Bailly, J.-L. Pépin, Factors Associated With Residual Apnea-Hypopnea Index Variability During CPAP Treatment, *Chest* 163 (2023) 1258–1265. doi:10.1016/j.chest.2022.12.048.
 - [42] M. Nissim, L. Abzianidze, K. Evang, R. Van Der Goot, H. Haagsma, B. Plank, M. Wieling, Sharing Is Caring: The Future of Shared Tasks, *Computational Linguistics* 43 (2017) 897–904. doi:10.1162/COLI_a_00304.
 - [43] M. Srikumar, R. Finlay, G. Abuhamad, C. Ashurst, R. Campbell, E. Campbell-Ratcliffe, H. Hongo, S. R. Jordan, J. Lindley, A. Ovadya, J. Pineau, Advancing ethics review practices in AI research, *Nat Mach Intell* 4 (2022) 1061–1064. doi:10.1038/s42256-022-00585-2.
 - [44] R. Schaeffer, J. Kazdan, Y. Denisov-Blanch, B. Miranda, M. Gerstgrasser, S. Zhang, A. Haupt, I. Gupta, E. Obbad, J. Dodge, J. Forde, F. Orabona, S. Koyejo, D. Donoho, Position: Machine learning conferences should establish a "refutations and critiques" track, in: *NeurIPS 2025*, volume 38, 2025.